DOI:
10.1080/08982112.2015.1100463

21st Century Screening Experiments: Discussion

Steven G. Gilmour
Southampton Statistical Sciences Research Institute
University of Southampton
Highfield
Southampton SO17 1BJ
UK

July 3, 2015

First, I thank the author for an interesting paper, which I hope will help to bring good methods of designing experiments to a broader audience. My comments come in two parts and are constructive, rather than critical (though the name "definitive screening designs" is easy to criticize as being rather pretentious). On the design side, I will show how these designs are really just special cases of a much broader class of designs chosen to have good estimation properties across a large set of possible models. On the analysis of the resulting data I will express a bit more skepticism, not specifically of the methods described by Jones, but of what the entire research community seems to be trying to do; I wonder if we are asking for too much?

1 What makes a good screening design - a broader class than definitive screening designs?

As explained by Jones, definitive screening designs (DSDs) have a number of properties which make them attractive for many screening experiments:

1. they have a small number of runs, order of the number of factors;
2. they allow orthogonal estimation of main effects;
3. main effect estimates are uncorrelated with two-factor interaction estimates;
4. two-factor interaction estimates are not completely confounded with each other;
5. they use three levels, allowing estimation of quadratic effects; and
6. they have good projective properties.

It is difficult to argue with any of these as desirable qualities a screening design should have. However, they need not be absolute restrictions, such that any design which nearly meets them is to be condemned. In the terminology of W. Müller, these are *hard* criteria. I would modify the list of desirable properties to involve more *soft* restrictions, as follows:

1. they should have a small number of runs, but with flexibility to choose more or fewer according the practical situation at hand;
2. main effects should be nearly orthogonally estimated, with small non-zero correlations being acceptable;
3. main effects should have low correlations with two-factor interactions, but again they do not have to be absolutely zero;
4. two-factor interactions should be confounded with each other as little as possible;
5. they should allow estimation of quadratic effects by using three levels, for some or all of the factors, as appropriate to the situation at hand; and
6. they should have good projective properties, not in a very vague sense, but specifically onto the dimensions that *a priori* are expected to be the right size.

A broader class of screening designs can be obtained, using various construction methods, by applying the Q_B class of optimality criteria of Tsai, Gilmour and Mead (2007). These criteria are obtained as rough approximations to weighted averages of A_S-optimality criteria over many possible models of interest. We start with a *maximal* model, containing all factorial and polynomial effects which might be of interest, written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is the vector of responses, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}' = [\beta_0 \cdots \beta_v]$ is the vector of parameters and $\boldsymbol{\epsilon}$ is a vector of random errors having $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The maximal model need not be estimable from the design, e.g. in the situations described in the paper, it could be the model consisting of main effects and two-factor interactions, as in Figures 1-5, or the full second order model, as in later sections of the paper.

The approximation proceeds as follows. Let a_{ij} be the elements of $\mathbf{X}'\mathbf{X}$. Ignoring higher order terms in the diagonal expansion of $|\mathbf{X}'\mathbf{X}|$ and in the Taylor series expansion of the inverse, we get

$$V\left(\hat{\beta}_i\right) \approx \sum_{j=0}^{v} \frac{a_{ij}^2}{a_{ii}^2 a_{jj}}.$$

The weighted-average approximate A_s-efficiency over all candidate models is

$$Q_B = \sum_{i=1}^{v}\sum_{j=0}^{v} \frac{a_{ij}^2}{a_{ii}^2 a_{jj}} p_{ij},$$

where p_{ij} is the sum, over models which include both β_i and β_j, of the prior probabilities of a model being the best (in this class of models. See Tsai, Gilmour and Mead (2000, 2007) for more details of this approximation. The approximation to the weighted average of A_S criterion functions can be so poor as to be useless for that purpose, but it is still useful for comparing designs. Tsai, Gilmour and Mead (2004) showed that the ordering of designs is very close to that which would be obtained by a true weighted average A_S-optimality.

The general form of criterion requires a very large number of prior probabilities to be specified. Under exchangeability of factor labels, however, we need only specify π_1, the probability of a linear effect being in the best model, π_2, the probability of a quadratic effect being in the best model, given the corresponding linear effect is in the best model, and π_3, the probability of an interaction being in the best model, given that both of the marginal linear effects are in the best model. Then we only have to count the candidate models to evaluate the criterion.

The set of Q_B criteria are very flexible. We can move smoothly from the hard criteria to (or towards) A_s-optimality by adjusting the priors, which act like hardness parameters, getting many different designs on the way. Tsai and Gilmour (2010) showed this in general and Tsai and Gilmour (2015) studied the case of saturated two-level main-effects designs in more detail. These criteria are even more flexible than this. We can nest soft criteria inside hard criteria, e.g. search for a Q_B-optimal among the class of orthogonal main effects designs, as in Tsai *et al.* (2000, 2004) for three-level designs. We can have any number of factors, with any number of levels, in any number of runs, with any maximal (linear) model; blocking can also be allowed for, as will be shown in forthcoming work.

With such flexibility, it is not surprising that Q_B is closely related to many other well-known criteria. Tsai and Gilmour (2010) showed that, with appropriate maximal models and priors, Q_B is equivalent to:

- $E(s^2)$ for supersaturated or saturated two-level designs;
- G_2-generalized aberration for two-level factors;
- generalized minimum aberration (GMA) for three-level qualitative factors;
- β-aberration for continuous factors; and
- weighted-A-optimality for regular fractions.

The proofs of these results are quite complex, but they are not essential. All that is needed to use Q_B in practice is:

1. a maximal model;
2. prior probabilities for effects of each type being in the best model; and
3. a method for finding optimal or near-optimal designs.

The best design construction method depends on the type of design being sought. A general coordinate- or point-exchange algorithm will usually do a reasonably good job, if there are no other restrictions. However, some of the problem-specific methods which have been used are the following.

- Tsai *et al.* (2000) used a columnwise-pairwise algorithm to find orthogonal and nearly-orthogonal three-level main effects designs. The restriction that main effects should be as nearly orthogonally estimated as possible means that each factor should be level-balanced, so that an algorithm which makes use of this restriction is appropriate.

- Tsai *et al.* (2000) also showed how fold-over methods could be used to obtain three-level screening designs, giving the same designs as described in Jones' paper, e.g. the 13-run design in Table 1 appears in Tsai *et al.* (2000).

- Further projection properties of these designs were used to rank them in a different way by Tsai *et al.* (2004).

- Tsai, Gilmour and Mead (2006) used Latin squares to construct further three-level main effects designs.

- Tsai and Gilmour (2015) used conference matrices to construct two-level saturated main effects designs.

2 Analysis of data from screening experiments

Jones stated that the purpose of screening experiments is to "identify the few important factors among many possible factors." The terminology can be clarified by defining a factor to be *important* if it is involved in at least one *active* effect. This is a property of the unknown parameters, not the outcome of a significance test using their estimates. What makes an effect active? Is it:

- having a non-zero coefficient;

- having a coefficient large enough to be of "practical importance"; or

- having a coefficient considerably larger than most of the other effects?

The answer to this question affects how we should study the properties of screening designs.

In *The Design of Experiments*, 5th edition (1949), R.A. Fisher wrote

> Replication ... serves two distinct purposes. It serves to diminish the error ... (However,) its main purpose is to supply an estimate of error by which the significance of ... comparisons can be judged.

Screening designs have no replication, therefore no formal (frequentist) inference is possible, unless we can assume *a priori* that some reduced model is true. Under this assumption, however, we are not really screening. Hence, screening designs allow only *exploratory data analysis*. We should try to identify the few factors which stand out as being involved in the largest effects. Smaller effects might also be of practical importance, but they can wait for follow-up studies.

The details of the analysis proposed by Jones depend on the very specific properties of definitive screening designs and so are not robust to bad values. The broad strategy is:

1. find factors with large linear main effects; and then

2. keep them and search for large second order effects involving these factors.

This aims to be inferential. An exploratory, non-inferential analysis would proceed as follows:

1. find the best allowable p-parameter model for $p = 1, 2, \ldots, n$ (all allowable subsets or a sequential method);

2. use residual sum of squares (or similar) to find the smallest p for which the model captures a large chunk of the variation, preferably with $p \leq n/2$; then

3. the factors involved in the model are declared important.

Note that, if the design is better for first order effects than second order effects, then the model will be biased towards finding first order effects active and this might cast some doubt on the conclusions drawn from Xiang *et al.* (2006). Note also that effect heredity is the belief that two-factor interactions are much more likely to be active if both main effects are active. This is a prior probability statement, partly supported by historical data (e.g. Xiang *et al.*, 2006). The strength of this prior belief depends on the specific application. There are situations where it will not be true, e.g. after a steepest ascent. Heredity should not be confused with the principle of *marginality*, which states that if a two-factor interaction is in the model, the main effects of both factors involved must be in the model, *whether or not these main effects are active*. This is a mathematical rule which is required for models to make sense. Marginality must always be used to ensure, for example, that the number of parameters in our model do not depend on the coding of factors. Any exceptions imply that the interpretation of parameters is not really in terms of main effects and interactions.

References

Tsai, P. W. and Gilmour, S. G. (2010). A general criterion for factorial designs under model uncertainty. *Technometrics* **52**, 231–242.

Tsai, P. W. and Gilmour, S. G. (2015). New families of Q_B-optimal saturated two-level main effects screening designs. *Statistica Sinica* in press.

Tsai, P. W., Gilmour, S. G. and Mead, R. (2000). Projective three-level main effects designs robust to model uncertainty. *Biometrika* **87**, 467–475.

Tsai, P. W., Gilmour, S. G. and Mead, R. (2004). Some new three-level orthogonal main effects plans robust to model uncertainty. *Statistica Sinica* **14**, 1075–1084.

Tsai, P. W., Gilmour, S. G. and Mead, R. (2006). Statistical isomorphism of three-level fractional factorial designs. *Utilitas Mathematica* **70**, 3–10.

Tsai, P. W., Gilmour, S. G. and Mead, R. (2007). Three-level main-effects designs exploiting prior information about model uncertainty. *Journal of Statistical Planning and Inference* **137**, 619–627.